SCVO: ADDRESSING SPARSE BUT CRITICAL VARIABLE OVERWHELM IN VLMS FOR ADVERTISING IMAGE PREFERENCE PREDICTION ACROSS MULTI-COUNTRY MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision language models (VLMs) have demonstrated remarkable capabilities in multimodal tasks, yet their sensitivity to sparse, critical, and overwhelmed variables remains unexplored. The image preference prediction across multi-country markets task serves as a representative case in this regard. Specifically, VLMs (e.g., QwenVL) are tasked with judging between two images (A and B) for the same product across diverse markets (e.g., Korea, France), the model's predictions often collapse to a single output (e.g., always "A") despite ground-truth preferences varying by country. This failure is attributed to Sparse Critical Variable Overwhelm (SCVO): the model is overwhelmed by dominant high-volume variables (e.g., product attributes, image patches consuming hundreds of tokens), while the critical low-volume variables (e.g., country names consuming only a few tokens) is statistically drowned out. To study this, we firstly collect dataset, a real-world advertising image click-through preference across multi-country markets, and then a novel training framework that strategically mtigate SCVO is presented and used to trained with the dataset yiedling to CountryReward, a judge model for advertising image preference prediction across multi-country markets. Our framework involves three tailored modules: (1) a cross-country retrieval augmentation generation that injects historical click-through preferences aligned with target markets into the model training, enhancing localized relevance prediction. (2) a country adapter module that dynamically modulates image representations based on textual country embeddings, enabling precise visual preference adaptation for diverse markets. (3) an focus-driven penalty loss function that penalizes mispredictions related to the overlooked variable more heavily. Finally, we apply the CountryReward as the reward model to fine-tune VLMs through Reinforcement Learning (RL) which can output background designs fed to text-to-image model (e.g., SDXL) and generate effective e-commerce image for targeted country. Experiments on a the proposed dataset show that our approach significantly mitigates the SCVO effect and improves the preference prediction accuracy. This work highlights the need for robust handling of sparse critical variables in VLMs and offers a scalable solution for real-world applications where subtle contextual shifts drive decision-making.

1 INTRODUCTION

Vision-language models (VLMs) (Wang et al., 2024; Chen et al., 2024) have emerged as a cornerstone of modern artificial intelligence, demonstrating remarkable proficiency across a broad spectrum of multimodal tasks, from visual question answering and image captioning to complex reasoning about visual scenes. Their ability to learn powerful multimodal representations have been used to be multimodal reward models (RMs) (He et al., 2024; Liu et al., 2025a; Xu et al., 2025), which can provide crucial reward signals to guide model training (Ouyang et al., 2022; Rafailov et al., 2024; Schulman et al., 2017) and inference (Gulcehre et al., 2023; Snell et al., 2024). However, despite their impressive performance, a critical aspect of their real-world applicability remains underex-

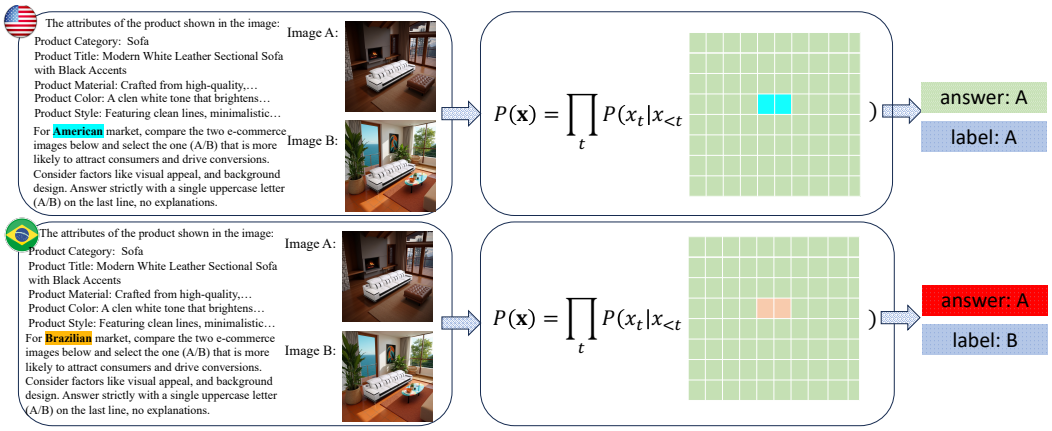


Figure 1: Sparse Critical Variable Overwhelm (SCVO): The only different variable (blue or orange squares) shown in the two examples is the country names, which sparse but critical. Other cues (green squares shown in the image) including a pair of images for the same product and product attributes are same, which are high-volume and dominated. The model's autoregressive decision-making process collapses, producing a market-invariant prediction.

plored: their robustness and sensitivity to specific, often underrepresented, variables in instruction-following scenarios. In particular, the sensitivity of VLMs to the sparse but critical variables which are overwhhelmed in excessive or complex input instruction, poses a significant challenge to their reliability and trustworthiness.

The image preference prediction across multi-country markets task is a representative case in the regard mentioned above. As shown in Figure 1, in this setting, models such as QwenVL (Bai et al., 2023) are required to discern nuanced preferences between two images (A and B) depicting the same product across distinct markets, for instance, comparing consumer choices in American market versus Brazilian market. Despite clear empirical evidence that human preferences vary significantly by region, VLMs frequently exhibit a collapse in their decision-making process, defaulting to a single output choice (e.g., persistently selecting "A") irrespective of the target market. This failure is attributed to Sparse Critical Variable Overwhelm (SCVO): a VLM's autogressive probability chain, $P(\mathbf{x}) = \prod_t P(x_t|x_{<t})$, is dominated by high-volume variables (e.g., product attributes, image patches consuming hundreds of tokens), causing the influence of the sparse critical variable (e.g., country names consuming only a few tokens) to be statistically drowned out during attention-weighted feature fusion. The green squares in Figure 1 represented as the same cues drowning out the influence of the blue or orange squares, resulting the model fails to allocate sufficient sensitivity to the critical variable, breaking chain rule dependence.

To study this, we first collect a dataset, called Multi-Country Ad Click Preference (MACP), a real-world advertising image click-through preference across multi-country markets. Our dataset contains 823K training samples and 18K test samples involving 10 countries. The number of sample from different countries is uniform. Each sample includes two different images of the same product, their Click-Through Rate (CTR), the detailed product information, including titles, categories, and other relevant attributes. These samples are collected from same e-commerce platform, ensuring reliability in data source and characteristics.

In the domain of advertising image preference prediction across different country markets, the presence of SCVO poses significant challenges to model accuracy. To address this issue, we introduce a novel training framework specifically designed to mitigate SCVO-related limitations. The proposed approach leverages a comprehensive dataset to train CountryReward, a specialized judge model for advertising image preference prediction across different country markets. Our framework incorporates three strategically designed components: (1) A cross-country retrieval augmentation generation that integrates historical click-through perferences aligned with target markets, thereby enhancing the model's capacity for localized relevance assessment. (2) A country adapter module that dynamically adjusts visual representations using textual country embeddings, enabling fine-grained adapta-

tion to diverse market-specific characteristics. (3) A focus-driven penalty loss function that assigns weighted penalties to prediction errors associated with previously overlooked variables. Through these innovations, our framework significantly mtigates the SCVO effect and improves prediction accuracy while maintaining robustness across varied market environments.

In the domain of cross-border e-commerce visual content generation, accurately aligning generated imagery with market-specific preferences remains a challenging task. To address this limitation, we integrate the CountryReward as a reward model to fine-tune VLMs through Reinforcement Learning (RL). The optimized VLM subsequently generates detailed background design tailored to specific markets, which serve as inputs to text-to-image models such as SDXL (Podell et al., 2023). The final output consists of highly targeted e-commerce images designed to resonate with consumers in particular countries, thereby enhancing visual relevance and commercial effectiveness.

We summarize our contributions as four aspects:

- Identifying and formalizing a novel research problem: This work is the first to systematically identify a critical deficiency in VLM, Sparse Critical Variable Overwhelm (SCVO).
- We collect the *Multi-Country Ad Click Preference (MACP)* dataset, a novel real-world e-commerce advertising image click-through preference data from 10 countries.
- We design an innovative training framework for the *CountryReward* proposed, a judge model can accurately predict image preference across multi-country markets. This framework integrates three tailored modules:
  - A *cross-country retrieval augmentation generation* that enhances the model's understanding of localized relevance by leveraging historical click-through data aligned with target markets.
  - A *country adapter module* that enables fine-grained adaptation to diverse market-specific features by dynamically adjusting visual representations using textual country embeddings.
  - A *focus-driven penalty loss* that can adaptively apply varying penalities based on focus of features such as country, image, and product when a prediction error occurs.
- We further use *CountryReward* as a reward model to fine-tune VLMs via Reinforcement Learning (RL), enabling the generation of country-market adapted background designs through text-to-image models (e.g., SDXL) for targeted e-commerce applications.

2 Related Work

2.1 Multimodal Reward Models

Multimodal reward models play an increasingly critical role in aligning vision understanding and generation systems with human preferences. A widely adopted strategy involves fine-tuning visual-language models (VLMs) (Li et al., 2024; Bai et al., 2022), capitalizing on their strong multimodal alignment capacities to acquire reward functions reflective of human judgments. Previous research has investigated reward modeling in the context of visual generation (Liu et al., 2025a; Xu et al., 2025; He et al., 2024; Wang et al., 2025b) and visual understanding tasks (Zang et al., 2025; Xiong et al., 2025). For example, Ziegler et al. (2020) devises an efficient pipeline for building multimodal preference datasets and utilizes existing high-quality data to train IXC-2.5-Reward, a model capable of accurately assessing outputs from visual understanding tasks. Similarly, Wang et al. (2025b) gathers human feedback to create a dataset of human-rated videos used to train LiFT-Critic, a reward model designed to evaluate how closely generated videos match human expectations. Wang et al. (2025c) proposes UnifiedReward, a unified reward model that can evaluate both image and video generation along with understanding tasks, showing that collaborative learning across various visual domains leads to significant synergistic improvements. Wang et al. (2025a) presented UNIFIEDREWARD-THINK, a unified multimodal reward model based on lengthy Chain-of-Thought reasoning, facilitating multi-dimensional long-chain reasoning for visual understanding and generation tasks. Despite their promising performance, existing reward models does not further consider the setting where instructions contian spares but critical variables, frequently leading to imprecise or untrustworthy reward signals. To this end, we propose CountryReward, a reward

model can simultaneously adaptively consider different semantic clues (e.g., country name, product attributes, and image features), especially the sparse but critical cues which are overwhlmed by dominant features in the input.

2.2 Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022; Luong et al., 2024; Ziegler et al., 2020; Ouyang et al., 2022; Jiao et al., 2025; Zhang et al., 2025; Ying et al., 2024; Yang et al., 2024; OpenAI et al., 2024; Shao et al., 2024; Hui et al., 2024) collects human feedback regarding model outputs. The feedback is then used to optimize generative model via reinforcement learning methods such as PPO (Schulman et al., 2017), DPO (Rafailov et al., 2024) and GRPO DeepSeek-AI et al. (2025). The RL applications for VLMs contain visual quality assessment (Li et al., 2025), visual perception and reasoning (Liu et al., 2025b), mitigating hallucinations (Sun et al., 2023; Yu et al., 2024a), and aligning models with human preferences (Yu et al., 2024b; Zhou et al., 2024). As for briding VLMs and T2I models, a classifier (Wu et al., 2023) trained on human-curated image choices, which can output human preference score used to adapt T2I model. Parrot (Lee et al., 2024) optimizes the prompt expansion and T2I model network together via a multi-reward RL approach for imrpoving image quality. CAIG (Chen et al., 2025) first explores the utilization of VLMs for generating advertising images via optimizing for CTR as the object. Through RL method, CTR reward model is used to fine-tune VLMs. The fine-tuned VLMs can generate background designs, which input to T2I models to generate image better align with user preferences. However, the limitation they referred to is that their reward model overlook the preferences of niche market segments, whose lack of personalization could result in suboptimal experiences for diverse user segments. Moreover, our work can better integrate user preferences across different country market. We use our CountryReward, trained to overcome SCVO, as a preference reward model for fine-tuning a VLM. This allows the generative model to product bacground designs optimized for specific country markets that cater to the needs and behaviors for global users.

3 Method

3.1 CountryReward

As shown in Figure 2, our proposed model, named CountryReward, is built upon the Qwen2VL framework, incorporating several key innovations to ease SCVO and improve performance. The overall architecture consists of a vision transformer for image feature extraction, a language model for text understanding, and a country adapter mechanism. Additionally, we introduce a focus-driven regularization technique to guide the model's focus toward critical tokens (e.g., country, product, and image tokens). In addition, before the training, there is a retrieval augmentation generation process to create augmented choice based on the experince knowledge.

3.1.1 Cross-Country Retrieval Augmentation Gneration

To enhance the model's capacity for localized relevance prediction, we propose a Cross-Country Retrieval Augmentation Gneration (CC-RAG) that incorporates historical click-through preferences aligned with target markets. This module enables the model to leverage domain-specific behavioral patterns from regional users, thereby improving the model's sensitivity to the country variable, the sparese but critical variable. Considering the efficiency, CC-RAG is applied before training CountryReward to obtain the augment choice $\hat{y}_{\text{aug}}$.

Given a query instance from country $c \in \{\text{US}, \text{FR}, \text{KR}, ...\}$ with text embedding $\mathbf{q}_t \in \mathbb{R}^d$ and candidate image embeddings $\{\mathbf{q}_i^A, \mathbf{q}_i^B\}$, we first retrieve the most relevant historical items through a two-stage hierarchical retrieval process: "Text-based Retrieval" and "Image-based Retrieval". The detailed process is as follows:

$$\mathcal{N}_t = \text{Top-}k\left(\mathbf{q}_t \cdot \mathbf{T}_c^T\right), \quad \mathcal{N}_i^A = \text{Top-}m\left(\mathbf{q}_i^A \cdot \mathbf{I}_{\mathcal{N}_t}^T\right), \quad \mathcal{N}_i^B = \text{Top-}m\left(\mathbf{q}_i^B \cdot \mathbf{I}_{\mathcal{N}_t}^T\right), \tag{1}$$

where $\mathbf{T}_c \in \mathbb{R}^{N \times d}$ represents the text embedding matrix for country c, $\mathcal{N}_t$ denotes the set of top-k semantically similar historical texts, $\mathbf{I}_{\mathcal{N}_t}$ contains image embeddings corresponding to $\mathcal{N}_t$, and $\mathcal{N}_i^A$ and $\mathcal{N}_i^B$ denote the set of top-k similar historical images.

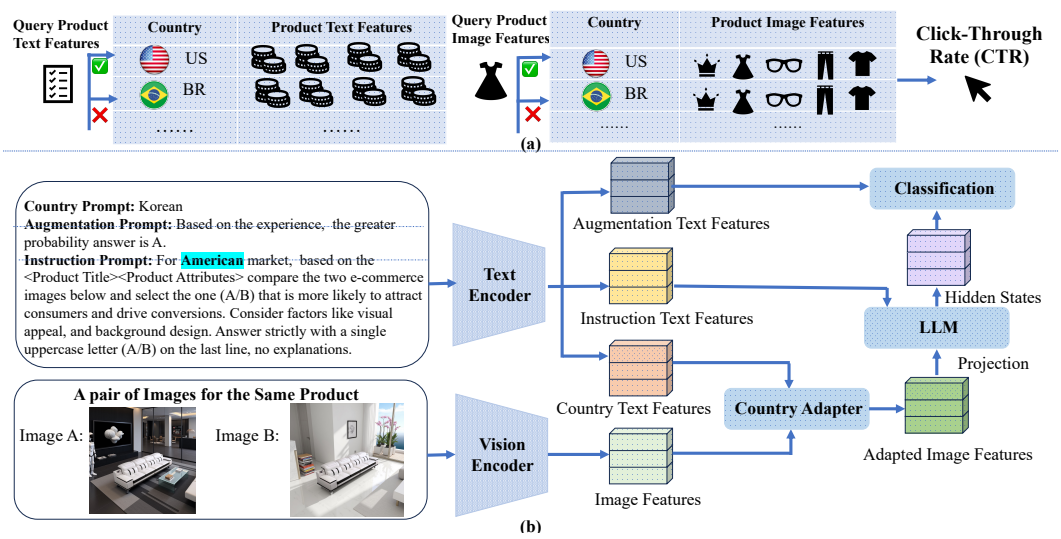


Figure 2: CountryReward: Figure (a) represents that based on the American knowledge database, the query product text features firstly retrieve items which have similar product text features, and then the query product image features retrieves the ctr values among the items retrieved in the frist stage. Figure (b) represents the training framework of CountryReward.

The retrieved items are aggregated using a position-aware weighting scheme that assigns higher importance to more relevant neighbors. For each retrieved item at position i, we assign weight $w_i = n - i$, where n is the total number of retrieved items. The preference scores for candidates A and B are computed as:

$$S_A = \frac{\sum_{i=1}^{n} w_i \cdot \mathbb{I}[\text{CTR}_A^{(i)} \geq \text{CTR}_B^{(i)}]}{\sum_{i=1}^{n} w_i}, \quad S_B = \frac{\sum_{i=1}^{n} w_i \cdot \mathbb{I}[\text{CTR}_B^{(i)} > \text{CTR}_A^{(i)}]}{\sum_{i=1}^{n} w_i} \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function, and $\text{CTR}_A^{(i)}$, $\text{CTR}_B^{(i)}$ represent the historical click-through rates of the i-th retrieved item, and the final augmented prediction is determined by:

$$\hat{y}_{\text{aug}} = \begin{cases} A & \text{if } S_A > S_B \\ B & \text{otherwise} \end{cases} \tag{3}$$

3.1.2 Country Adapter Module

Effectively adapting large vision-language (VL) models to diverse global markets requires sensitivity to country-specific, visual content preferences. To this end, we introduce a Country Adapter Module (CAM). Inspire by FiLM (Perez et al., 2017), this module dynamically modulates the visual features extracted by the vision encoder based on textual embeddings derived from country-specific information, allowing the model to adjust its perceptual processing for different country markets.

The core mechanism involves generating a set of affine transformation parameters (scale and shift) from a learned country embedding. Let $\mathbf{c}_i \in \mathbb{R}^d$ denote the mean-pooled embedding vector of the tokenized country name for the i-th sample in a batch, where d is the hidden dimension size.

The adaptation parameters are generated by a small feed-forward network, the Country Adapter:

$$\gamma_i, \beta_i = \text{Split}(\text{CountryAdapter}(\mathbf{c}_i)) \tag{4}$$

where CountryAdapter: $\mathbb{R}^d \to \mathbb{R}^{2d}$ is implemented as:

$$\text{CountryAdapter}(\mathbf{c}_i) = \mathbf{W}_2(\text{ReLU}(\mathbf{W}_1\mathbf{c}_i + \mathbf{b}_1)) + \mathbf{b}_2 \tag{5}$$

Here, $\mathbf{W}_1 \in \mathbb{R}^{d/2 \times d}$, $\mathbf{b}_1 \in \mathbb{R}^{d/2}$, $\mathbf{W}_2 \in \mathbb{R}^{2d \times d/2}$, $\mathbf{b}_2 \in \mathbb{R}^{2d}$ are learnable parameters. The output is split into two vectors $\gamma_i \in \mathbb{R}^d$ (scale) and $\beta_i \in \mathbb{R}^d$ (shift).

Let $\mathbf{V}_i \in \mathbb{R}^{N \times d}$ represent the sequence of visual features (e.g., N image patch embeddings) corresponding to the i-th sample before integration into the language model's input embedding space. The adapted visual features $\tilde{\mathbf{V}}_i$ are computed via an element-wise affine transformation:

$$\tilde{\mathbf{V}}_i = \gamma_i \odot \mathbf{V}_i + \beta_i \tag{6}$$

where $\odot$ denotes the Hadamard (element-wise) product. This transformation is applied to the entire set of visual features $\mathbf{V}_i$ associated with the specific country embedding $\mathbf{c}_i$. This allows the model to selectively emphasize or suppress certain visual patterns based on learned country-specific cues, effectively tailoring the visual representation to relevant country markets.

3.1.3 Focus-Driven Penality Loss

To enhance the model's ability to leverage multimodal inputs effectively and improve country-specific adaptation, we propose a novel focus-driven penalty loss (FDPL), which is designed to penalize the model when it fails to adequately attend to input components (e.g., country tokens, product descriptors, or image features) during erroneous predictions, while imposing no additional penalty for correct predictions. This is achieved by introducing an auxiliary penalty term that is dynamically scaled based on the relative focus allocated to each key component. Let $\mathbf{H} \in \mathbb{R}^{T \times d}$ denote the hidden states of the final transformer layer, where T is the sequence length and d is the hidden dimension. The hidden states $\mathbf{H}$ are obtained by feeding the adapted visual features $\tilde{\mathbf{V}}$ (stated in 3.1.2) and instruction text features $\tilde{\mathbf{T}}$ into the VLM. Next, for each sample in a batch, we identify the token positions of key input components: country token t_c, product token t_p, and image token t_i. The focus intensity toward each component is approximated using the L2-norm of their corresponding hidden states:

$$\text{Focus}_c = \frac{\|\mathbf{H}[t_c]\|_2}{\sum_{j=1}^{T} \|\mathbf{H}[j]\|_2}, \quad \text{Focus}_p = \frac{\|\mathbf{H}[t_p]\|_2}{\sum_{j=1}^{T} \|\mathbf{H}[j]\|_2}, \quad \text{Focus}_i = \frac{\|\mathbf{H}[t_i]\|_2}{\sum_{j=1}^{T} \|\mathbf{H}[j]\|_2}, \tag{7}$$

where $\|\cdot\|_2$ is the L2-norm. The penalty terms for country, product, and image are defined as:

$$\mathcal{P}_c = 1 - \text{Focus}_c, \quad \mathcal{P}_p = 1 - \text{Focus}_p, \quad \mathcal{P}_i = 1 - \text{Focus}_i. \tag{8}$$

These penalties are activated only when the model makes an incorrect prediction. For a batch of size B, let $\hat{y}_i$ and y_i be the predicted probability and ground truth label for the i-th sample, respectively. The indicator function for incorrect prediction is:

$$\mathbb{I}_i = \begin{cases} 1 & \text{if } (\hat{y}_i \geq 0.5) \neq (y_i = 1), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The total penalty loss for the batch is computed as:

$$\mathcal{L}_{\text{penalty}} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}_i \cdot \left(\mathcal{P}_c^{(i)} + \mathcal{P}_p^{(i)} + \mathcal{P}_i^{(i)} \right). \tag{10}$$

The overall training objective combines the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ with the penalty loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\sigma(\mathbf{W}_{\text{classifier}} \cdot (\mathbf{e}_{\text{aug}} + \mathbf{h}_{\text{last}})), y) + \lambda \mathcal{L}_{\text{penalty}}, \tag{11}$$

where following common practice in sequence classification with LLMs (Touvron et al., 2023), $\mathbf{h}_{\text{last}}$ is the last token of the hidden state as the discriminative representation, as it summarizes the contextual information of the entire input sequence, $\mathbf{e}_{\text{aug}}$ is the augmented features via extracting text embedding based on $\hat{y}_{\text{aug}}$ (stated in 3.1.1), $\mathbf{W}_{\text{classifier}}$ refers to the weight matrices in the classifier component of our model, σ is the sigmoid activation function, λ is a scaling hyperparameter (set to 0.1). This design encourages the model to strengthen its focus on under-attended components when errors occur, thereby improving feature utilization and country-specific decision-making.

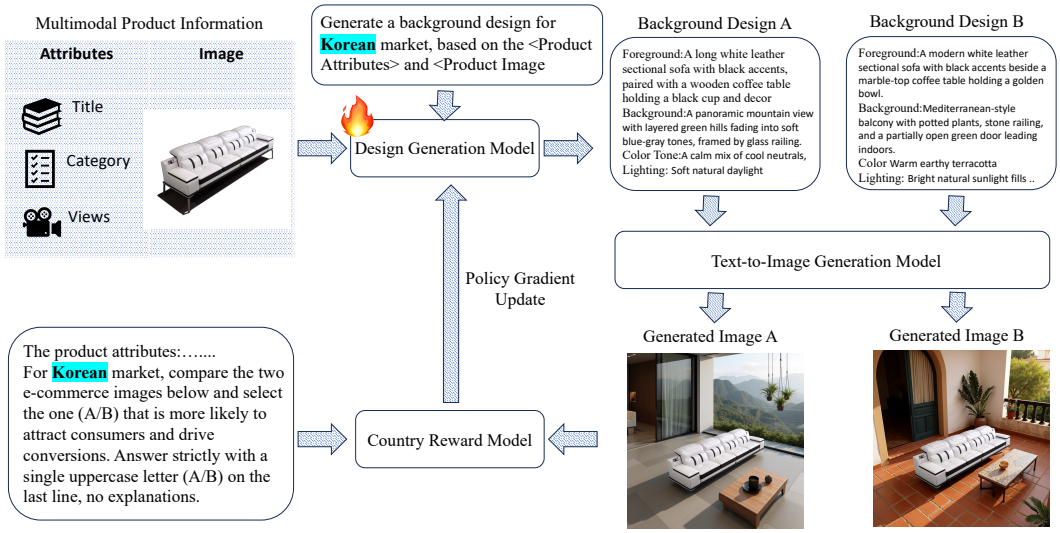


Figure 3: Country-Adapted Background Design Generation Framework: The Design Generation Model generates background design for target country, the T2I Generation Model creates product images according to the designs, the Country Reward Model then predicts the image preference according to the target country, giving feedback to optimize the Design Generation Model.

3.2 Country-Adapted Background Design Generation

To address the challenge of generating market-adapted visual content for cross-border e-commerce, we propose a reinforcement learning-based framework that leverages a CountryReward model to optimize the generation of country-specific background designs. The framework consists of three stages (shown in Figure 3):

Firstly, a Design Generation Model (DGM), implemented as a fine-tuned Qwen2-VL (Wang et al., 2024), a Vision-Language Model (VLM) trained on the proposed dataset, which generates textual background designs for the target country. The process of training and inference of DGM is:

$$d = \text{DGM}(\text{country}, \text{pro}, \text{I}_{ori}) \tag{12}$$

where $country$, pro, I_{ori} represent the target country, the product attributes, and the original product image respectively. Next, the pair of generated background designs d_A and d_B will be obtained via $d_A = \text{DGM}(country, pro, I_{ori})$ and $d_B = \text{DGM}(country, pro, I_{ori})$.

Secondly, a controlled Text-to-Image (T2I) Generation Model, implemented as integrating a Stable Diffusion Model (Podell et al., 2023) with a ControlNet adapter (Zhang et al., 2023), that allows us to condition the generation process on a control map based on the canny edge of the product image. The component can enable the generated image not only aligns with target market preferences but also adheres to original product layout. The $\text{T2I}(p, I_{ori})$ function can be represented as follows:

$$\begin{aligned} z_{t-1} &= \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, \tau p, \text{Canny}(I_{ori}))) + \sigma_t \epsilon, \\ I &= \text{Decoder}(z_0), \end{aligned} \tag{13}$$

where z_t is the latent representation at timestep t, I_{ori} is the input image, p is the text prompt, τp is the text encoder, αt, $\bar{\alpha}_t$, σ_t are the noise scheduling parameters, $\text{Canny}(\cdot)$ is the canny edge extraction function, and $\text{Decoder}(\cdot)$ can decode the final latent z_0 to the generated image I. We will obtain the pair of generated images I_A and I_B via $I_A = \text{T2I}(d_A, I_{ori})$ and $I_B = \text{T2I}(d_B, I_{ori})$.

Thirdly, a Country Reward Model (NRM) that predicts the country-specific preference for the pair of generated images (I_A and I_B). According to the obtained preference choice, the design of more attractive image is denoted as d^+, and the design of the less attractive image is represented as d^-. In order to fine-tune the DFM to choose higher attractive design d^+ and reject less attractive ones

Table 1: Comparison of accuracy performance across different reward models on MACP. Both accuracy and sensitivity perform better with higher values, and their units are percentage (%).

Model	Accuracy	Sensitivity	BR	CL	ES	FR	KR	JP	US	MX	AU	SA
SAIL2-8B	49.26	26.32	48.96	48.85	49.35	49.38	49.05	49.12	49.15	49.95	49.53	49.28
InternVL3-8B	49.15	27.10	49.86	48.67	50.04	49.04	49.99	48.25	48.94	49.48	48.42	48.54
Qwen2-VL-7B	49.76	31.55	49.77	50.26	48.85	49.16	48.20	50.58	51.17	49.18	49.75	50.70
Qwen2-VL-7B (finetuned)	44.61	20.82	46.96	43.71	49.76	46.15	47.78	40.99	42.99	46.66	40.92	40.16
Qwen2-VL-7B (with FC Head)	55.60	36.73	53.69	56.57	50.86	54.41	53.07	59.10	56.94	53.75	58.47	59.21
CountryReward (w/o CC-RAG)	56.81	37.40	54.93	57.05	53.80	55.39	54.48	59.57	57.65	55.64	59.52	60.11
CountryReward (w/o CAM)	57.98	38.95	55.50	57.85	53.78	56.75	55.19	61.47	59.39	55.98	61.71	62.21
CountryReward (w/o FDPL)	56.95	37.47	54.79	57.09	52.80	56.12	54.44	61.01	58.79	55.33	58.83	60.33
CountryReward	**60.37**	**40.84**	**58.70**	**61.12**	**56.33**	**59.38**	**57.88**	**64.54**	**61.82**	**58.72**	**63.30**	**62.93**

d^-. The feedback signals provided by NRM to refine the DGM via Direct Preference Optimization (DPO) (Rafailov et al., 2024). Speicifically, given an optimization policy model DGM_θ and a reference model DGM_{ref}, the optimization object is:

$$\mathcal{L}_{dpo} = -log\sigma(\beta log\frac{\text{DGM}_\theta(d^+|\text{country, pro, I}_{ori})}{\text{DGM}_{ref}(d^+|\text{country, pro, I}_{ori})} - \beta log\frac{\text{DGM}_\theta(d^-|\text{country, pro, I}_{ori})}{\text{DGM}_{ref}(d^-|\text{country, pro, I}_{ori})}), \tag{14}$$

where σ and β are the sigmoid activation function and a regularization parameter respectively. DGM_θ and DGM_{ref} are policy and reference models repectively, where the policy one is ooptimized while the reference one is frozen. In addition, we utilize the fine-tuned DGM to generate background designs for products. These designs are then fed into the T2I Generation Model to create product advertising images ensuring that the generated background designs are tailored to the target country's preferences.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate our proposed method on the collected Multi-Country Ad Click Preference (MACP) dataset. The dataset comprises 823K training samples and 180K test samples, uniformly distributed across 10 distinct country markets, including "BR", "CL", "ES", "FR", "JP", "US", "MX", "AU", and "SA". Each sample contains detailed product information, including titles, categories, tags, and other relevant attributes, two different advertising images (A and B) for the same product, and the Click-Through Rate (CTR) indicating user preference in the specific market. The dataset is sourced from a major cross-border e-commerce platform, containing 67K product samples with 250K unique advertising images, and ensuring consistency in data source and characteristics.

4.2 Analysis on CountryReward

Evaluation Metric. To evaluate the performance of our CountryReward, we introduce the Accuracy and Sensitivity metrics. Accuracy measures the proportion of correct predictions, and Sensitivity measures the proportion of simultaneous correct predictions across different country combinations, reflecting the cross-country consistency sensitivity of the model's predictions, which are defined as:

$$\text{Accuracy} = \frac{1}{B}\sum_{i=1}^{B}\mathbb{I}(\hat{y}_i = y_i), \quad \text{Sensitivity} = \frac{\sum_{i=1}^{N}\sum_{(c_j,c_k)\in\mathcal{C}_2(S_i)}\mathbb{I}(\hat{y}_{i,c_j} = y_{i,c_j} \wedge \hat{y}_{i,c_k} = y_{i,c_k})}{\sum_{i=1}^{N}|\mathcal{C}_2(S_i)|} \tag{15}$$

where N represents the total number of samples, $\hat{y}_i$ denotes the predicted class label for the i-th sample, obtained by thresholding the sigmoid normalized logits at 0.5 and mapping to class labels A, B, y_i corresponds to the ground-truth label, N is represented as total number of unique items, S_i is the set of countries for item i, $\mathcal{C}_2(S_i)$ is the set of all 2-combinations of countries in S_i, $\hat{y}_{i,c}$ is the predicted answer for item i in country c, $y_{i,c}$ is the ground truth answer for item i in country c, and $\mathbb{I}[\cdot]$ is the indicator function (1 if condition true, 0 otherwise).

Table 2: Comparison of performance across different DGM on MACP. The unit of CountryReward is percentage (%).

Model	Metric	Accuracy	BR	CL	ES	FR	KR	JP	US	MX	AU	SA
DGM (w/o RL)	CountryReward	56.04	54.03	57.05	51.67	55.40	53.59	60.40	57.64	54.36	56.38	59.89
DGM	CountryReward	**59.60**	**58.36**	**60.10**	**54.66**	**57.31**	**56.22**	**61.76**	**60.28**	**57.33**	**68.26**	**61.71**

Quantitative Results. As shown in Table 1, experimental results on our MACP benchmark demonstrate that SAIL2-8B (Yin et al., 2025), Internvl3-8B (Zhu et al., 2025), and Qwen2-VL-7B (Wang et al., 2024) exhibit a significant performance gap, with accuracy approximately 11.11%, 11.23%, and 10.61% lower than our proposed method (60.37%) respectively, alongside notably poorer sensitivity, and when the original Qwen2-VL model is fine-tuned on the MACP dataset using a standard approach, the model exhibited a complete prediction collapse. These performance degradations primarily stem from their vulnerability to SCVO effect.

Ablation Study. To dissect the contribution of each proposed component, we conduct ablation studies on the test set. The results are summarized in Table 1. Removing the Cross-Country Retrieval Augmentation Generation (w/o CC-RAG) leads to a 3.56% drop in overall accuracy and 3.44% drop in sensitivity. This highlights the importance of injecting historical market-specific preference knowledge to guide the model. Removing the Country Adapter Module (w/o CAM) causes a more substantial drop of 2.39% in accuracy and drop of 1.89%. This underscores the critical role of dynamically modulating visual features based on country embeddings for adapting to local visual preferences. Removing the Focus-Driven Penalty Loss (w/o FDPL) results in a 3.42% accuracy decrease and in a 3.37% sensitivity decrease. This demonstrates that explicitly penalizing the model for under-attending to critical tokens during errors is an effective regularization strategy. The cumulative effect of all three modules is clear, as their removal (CountryReward-w/o-Modules) results in a significantly lower accuracy (55.60%) and lower sensitivity (36.73%).

Performance per Country. Table 1 shows the accuracy breakdown for each country. CountryReward achieves more balanced and higher performance across all countries compared to baselines. The variances of Qwen2VL-7B with FC Head and CountryReward are 7.12% and 8.18% respectively. CountryReward's specialized components obtain more robust adaptation to diverse markets.

4.3 Analysis on Country-Adapted Background Design Generation

CountryReward Evaluation. We use CountryReward to evaluate the quality of images generated using the optimized DGM versus backgrounds from the DGM without RL. As shown in Table 2, images generated using our method achieve a substantially higher CountryReward Score across all tested countries. This indicates that the optimized DGM produces background designs that lead to images better aligned with country-specific preferences.

Case Study. Figure 4 in appendix presents a case study for two products for five targeted country markets. This qualitative analysis demonstrates our method's capability to produce highly customized visual content that aligns with the preferences of diverse global markets. These results demonstrate our model's ability to capture nuanced, country-specific visual preferences, validating its effectiveness in mitigating SCVO and enabling tailored content generation for global markets.

5 Conculsion

This work identifies and addresses the Sparse Critical Variable Overwhelm (SCVO) problem in VLMs, where models fail to respond to instruction-critical variables that are sparse in the input space. We propose a novel training framework that effectively mitigates SCVO through integrated components including retrieval augmentation, a country adapter module, and a focus-driven penalty loss. Evaluated on the newly introduced MACP dataset, our resulting CountryReward model demonstrates significant improvements in cross-country preference prediction accuracy. Furthermore, we showcase its practical utility by employing it as a reward signal to optimize background design generation for targeted markets. This study provides a foundation for enhancing sensitivity to critical but sparse variables in multimodal reward models.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, Yuanjie Shao, Xinge You, Changxin Gao, and Nong Sang. Ctr-driven advertising image generation with multimodal large language models, 2025. URL https://arxiv.org/abs/2502.06823.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. URL https://arxiv.org/abs/2308.08998.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu,

Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation, 2024. URL https://arxiv.org/abs/2406.15252.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F. Chen, Shafiq Joty, and Furu Wei. Preference optimization for reasoning with pseudo feedback, 2025. URL https://arxiv.org/abs/2411.16345.

Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, Gang Li, Sangpil Kim, Irfan Essa, and Feng Yang. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation, 2024. URL https://arxiv.org/abs/2401.05675.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning, 2025. URL https://arxiv.org/abs/2503.22679.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback, 2025a. URL https://arxiv.org/abs/2501.13918.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025b. URL https://arxiv.org/abs/2503.01785.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning, 2024. URL https://arxiv.org/abs/2401.08967.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam

Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning, 2025a. URL https://arxiv.org/abs/2505.03318.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment, 2025b. URL https://arxiv.org/abs/2412.04814.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation, 2025c. URL https://arxiv.org/abs/2503.05236.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference, 2023. URL https://arxiv.org/abs/2303.14420.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models, 2025. URL https://arxiv.org/abs/2410.02712.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2025. URL https://arxiv.org/abs/2412.21059.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, Wenzhuo Liu, Xiao Liang, Shuicheng Yan, and Chao Feng. Sail-vl2 technical report, 2025. URL https://arxiv.org/abs/2509.14033.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning, 2024. URL https://arxiv.org/abs/2402.06332.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024a. URL https://arxiv.org/abs/2312.00849.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2024b. URL https://arxiv.org/abs/2405.17220.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model, 2025. URL https://arxiv.org/abs/2501.12368.

Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, and Zhi Jin. Codedpo: Aligning code models with self generated and verified source code, 2025. URL https://arxiv.org/abs/2410.05605.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. URL https://arxiv.org/abs/2402.11411.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A ETHICS STATEMENT

We confirm that this work adheres to the ethical guidelines and principles outlined in the ICLR 2026 Code of Conduct. We have conducted a careful review of the potential societal impacts of our work. To the best of our knowledge, we do not foresee our research being directly used for malicious purposes or contributing to significant negative societal consequences. All data used in this study are with applicable legal and ethical standards. We are committed to conducting research in a responsible and ethical manner and will continue to monitor the implications of our work.

B REPRODUCIBILITY STATEMENT

We confirm that the methodology presented in this paper is fully reproducible. To support transparency and facilitate further research, we will publicly release all data and source code used in our experiments upon acceptance of the paper. The code repository includes detailed instructions for environment setup, training, and evaluation to ensure easy replication of our results.

C LLM DISCLAIMER

We acknowledge the use of Large Language Models (LLMs) in the preparation of this manuscript. Specifically, Deepseek DeepSeek-AI et al. (2025) was used solely for two purposes: (1) to assist in literature review by summarizing existing research and identifying relevant papers, and (2) to polish the text for improved fluency and readability. All ideation, theoretical development, experimental design, data analysis, and result interpretation were conducted solely by the authors. The authors take full responsibility for the content, accuracy, and originality of the work presented herein.

D COUNTRY NAME ABBREVIATION

Table 3: Country name abbreviation

Abbreviation	**Full Name**
BR	The Federative Republic of Brazil
CL	Republic of Chile
ES	The Kingdom of Spain
FR	The French Republic
KR	Republic of Korea
JP	Japan
US	The United States of America
MX	The United Mexican States
AU	The Commonwealth of Australia
SA	Kingddom of Saudi Arabia

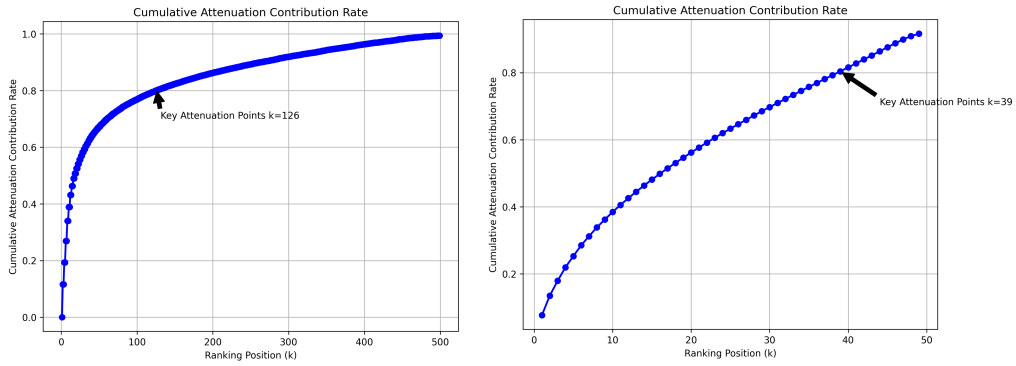
E CASE STUDY

This case study investigates product advertising adaptation across multicountry market by generating location-specific marketing imagery for two products. A compact blue car and a pair of white sneakers, across five distinct countries: France (FR), Korea (KR), Brazil (BR), Spain (ES), and the United States (US). This qualitative analysis demonstrates our framework's capability to produce highly customized visual content that aligns with the nuanced aesthetic preferences of diverse global markets, such as Parisian architecture for FR, traditional wooden interiors for KR, tropical coastal vistas for BR, Mediterranean urban textures for ES, and iconic desert or coastal landscapes for US. The study highlights the role of multimodal generative AI in scalable, location-aware marketing design, paving the way for automated, globally distributed visual campaigns that remain sensitive to regional identity and consumer expectations.

Figure 4: Case Study for two products across five distinct countries.

F Determination of the optimal K-Value in CC-RAG

To determine the optimal k value for top-k retrieval in our CC-RAG system, we employ a decay analysis method based on the cumulative attenuation contribution rate. Specifically, we compute the average cumulative decay of similarity scores across ranking positions from a large-scale retrieval experiment. The k value is set at the point where the cumulative decay contribution rate exceeds a threshold of 80%, indicating that including more results beyond this point yields diminishing returns. This data-driven approach ensures that we capture the majority of relevant information while maintaining efficiency.



(a) Top-K of Text Retrieval Augmentation Generation

(b) Top-K of Image Retrieval Augmentation Generation

Figure 5: The determination of the optimal K-values.

G Implementation Details

For CountryReward, we employ the Qwen2-VL-7B (Wang et al., 2024) as our foundation VLMs. In CC-RAG process, the values of top-k is 127 and 39 in text and image retrieval stage respectively. This training phase takes about 20 hours to complete. All experiments are conducted on a machine equipped with 8 NVIDIA A100 GPUs. To optimize training performance, DeepSpeed and FlashAttention-2 are adopted. We use a per-device batch size of 8, gradient accumulation steps of 2, learning rates of 1e-5, 5e-6, 1e-4, 1e-4 for projecter, LLM and country-adapter, classification, respectively, a cosine learning rate schedule, and 3 epochs with BF16 mixed-precision enabled. The λ is set to 0.1 in FDPL. For our T2I generation model, we use Stable Diffusion XL (Podell et al., 2023), enhanced with ControlNet (Zhang et al., 2023).

H AUGMENTATION STRATEGY IN CC-RAG

Table 4: Comparison of accuracy performance across different augmentation strategies on MACP. Both accuracy and sensitivity perform better with higher values, and their units are percentage (%).

Model	Accuracy	Sensitivity	BR	CL	ES	FR	KR	JP	US	MX	AU	SA
Qwen2-VL-7B (with FC Head)	55.60	36.73	53.69	56.57	50.86	54.41	53.07	59.10	56.94	53.75	58.47	59.21
Qwen2-VL-7B (with Instruct RAG)	54.58	36.49	52.41	54.10	50.34	53.76	51.86	57.81	55.74	52.51	58.36	58.94
Qwen2-VL-7B (with Embedding RAG)	55.69	37.23	53.54	56.70	51.45	54.37	52.98	58.68	57.20	53.97	**58.99**	58.98
Qwen2-VL-7B (with Scaled Embedding RAG)	**56.97**	**38.87**	**53.79**	**57.28**	**52.58**	**55.22**	**54.52**	**60.13**	**57.73**	**54.39**	56.05	**59.95**

Our investigation focuses on the effective incorporation of augmented answers (from Figure 2(a)) into CountryReward. We evaluate two paradigms (Table 4): instruction-based injection ("Qwen2-VL-7B with Instruct RAG") and embedding-based addition ("Qwen2-VL-7B with Embedding RAG") of the answer features to the discriminative features. Since both methods yielded inferior results to the baseline, we subsequently scaled the augmented text features to mitigate potential magnitude mismatches with the discriminative features.

$$\mathbf{E}_{\text{cls}} = \mathbf{h}_{\text{dis}} + \left(\frac{\|\mathbf{h}_{\text{dis}}\|_2}{\|\mathbf{e}_{\text{aug}}\|_2} \right) \cdot \mathbf{e}_{\text{aug}} \tag{16}$$

where $\mathbf{E}_{\text{cls}}$ is used to feed into classification head, $\mathbf{e}_{\text{aug}}$ is the extracted text embedding of augmented answers, and $\mathbf{h}_{\text{dis}}$ is the hidden states of the last token from the VLM.